回帰木を用いた NS/NNS テキスト分類

小林 雄一郎　（大阪大学）
kobayashi0721@gmail.com

1. はじめに

母語話者 (Native Speakers, NS) は，どうして非母語話者 (Non-native Speakers, NNS) の英語が母語話者によるものではないと分かるのであろうか。杉浦 (2004) が指摘するように，TOEFL でほぼ満点を取るような非母語話者であったとしても，発話された音声を聞くとすぐに非母語話者であることが母語話者に悟られてしまう。また，発音が関係のない作文であっても，非母語話者の作文は一目で母語話者に見破られてしまう。一体それは何故であろうか。日本の英語教育では語彙や文法は重要視されてきたため，大学生レベルの作文であればスペリングや文法に関する稚拙な誤りは比較的少ない。では，母語話者の作文と非母語話者の作文との差は何か。後者に見られる「不自然さ」とは何か。それらを部分的にでも明らかにしていくことは，英語教育において非常に重要な課題の 1 つである。

田中ほか (2006) は，母語話者と非母語話者による英語科学技術論文をデータとし，品詞 n-gram 分布に基づく NS/NNS 論文分類モデルを用いて，非母語話者の作文に存在する「非母語話者性」（統語的には誤りではないが，不自然さに強く関連する品詞レベルの要因）を抽出している。

だが，前述のように，母語話者の作文と非母語話者の作文との差は，必ずしも品詞レベルの要因だけではない。また，個々の要因の差異だけではなく，多くの要因が複雑に絡み合って，作用した結果として，後者に見られる「不自然さ」が生み出されている可能性も高い。そこで本研究の目的は，母語話者と非母語話者の大学生による論説文をデータとし，談話分析の観点から metadiscourse markers を説明変数に用いて，回帰木で母語話者の作文と非母語話者の作文を分類し，2 つの群を識別する特徴を抽出することにある。

2. データと方法論

2.1 データ

本研究で用いるデータは，日本人大学生の英作文データである ICLE-JP (the Japanese Component of International Corpus of Learner English) とアメリカ人大学生の英作文データ LOCNESS (the Louvain Corpus of Native English Essays) からそれぞれ 100 名分（計 200 名分）である。以下，前者を NNS，後者を NS とする。表 1 は，分析に用いるデータの概要である。

表 1: 使用データ

	N	Tokens	Types
NNS	100	48323	3718
NS	100	87950	7714

2.2 回帰木

回帰木とは，量的変数を扱う決定木であり，一定の規則（アルゴリズム）によって自動的にデータを分類していくモデルである。具体的には，対象データ全体を最もよく分類できる説明変数を探索し，それに従って分類されたデータ群に対しても，それぞれに最も分類効率の高い説明変数を探索するという作業を繰り返し，分類できなくなるまで分岐を行なう。そして，その分岐の繰り返しを樹形図として視覚化する。著名なアルゴリズムに C5.0, CART, CHAID などがあるが，本研究では CART (Classification and regression tree) を用いる。CART では，分岐の評価基準として，通常ジニ係数 (Gini index) を用いる。

2.3 Metadiscourse markers (MDM)

本研究における回帰木に用いる説明変数は，metadiscourse markers (MDM) である。MDM は，いくつかの相違点はあるものの，広義での談話標識の一種である。

MDM の研究において，最もよく使われる枠組みは，恐らく Hyland list (Hyland 2005) であろう。このリストは，様々な先行研究をベースとして，10 種類のカテゴリー（表 2）に分類される約 400 種類の談話表現を網羅的に収録し

たものである。また，このリストは，コーパスに基づく統計的研究を想定して作成されたものであり，これまでにアカデミック・ライティングを始め，教科書，学位論文，ビジネスレターなど，様々な言語データの分析で成果を上げている。

表 2: metadiscourse markers のカテゴリー

Category	*Examples*
Transitions (TRA)	but / because / in addition
Frame markers (FRM)	finally / to conclude
Endophoric markers (END)	notes above / see Fig
Evidentials (EVI)	according to X / (Y, 1990)
Code glosses (COD)	such as / in other words
Hedges (HED)	might / perhaps / possible
Boosters (BOO)	definitely / it is clear that
Attitude markers (ATM)	unfortunately / surprisingly
Engagement markers (ENG)	consider / you can see that
Self-mentions (SEM)	I / we / my / our

本研究では，個々のテキストにおける各カテゴリーの相対頻度を説明変数とし，NNS か NS かという群情報を目的変数として，回帰木を行なう。

そして，回帰木での分類に用いられた説明変数（カテゴリー）は，NNS と NS を識別する強い指標であるということを意味する。本研究では，これらのカテゴリーの使用例を量的・質的に分析する。

3. 結果と考察

3.1 回帰木の分類結果

図 1 は，回帰木による分類結果（クローズド・テスト）である。この図を見ると，先ず SEM の値（相対頻度）に基づいた分類が行なわれている。そして，SEM の値が一定以上であったデータに対しては，次に FRM の値に基づいた分類が行なわれる。また，SEM の値が一定未満であったデータに対しては，BOO の値に基づいた分類が行なわれる。さらに，BOO の値が一定以上であったデータでは再び SEM の値を参照し，逆に BOO の値が一定未満であったデータでは HED の値を参照する。

表 3 は，データを 2 分割して，一方を学習データ，他方を実験データとした予測を行なった結果（オープン・テスト）である。分類精度は，NNS テキストで 90.0%，NS テキストで 94.0%，全体では 92.0%であった。

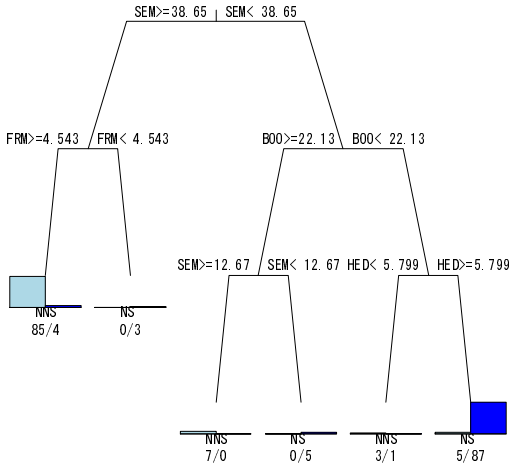

図 1: 回帰木による分類結果（クローズド・テスト）

表 3: 回帰木による分類結果（オープン・テスト）

	NNS	NS	correct ratio
NNS	45	5	90.0%
NS	3	47	94.0%
total			92.0%

3.2 分類に寄与した変数の分析

前節までの分析で，4 つの MDM カテゴリー (SEM, BOO, HED, FRM) を説明変数として，クローズド・テストで 95.0%，オープン・テストで 92.0%の精度で NNS テキストと NS テキストが分類されることが確認された。このことは，この 4 つのカテゴリーが両者を識別する強い指標であるということを意味する。以下，これらのカテゴリーの使用例を量的・質的に分析し，NNS テキストと NS テキストの特徴に光を当てていく。

3.3.1 Self-mentions (SEM)

SEM とは，書き手に対する明示的言及を指し，主に 1 人称の代名詞 (e.g. *I, we*) のことである。表 4 は，NNS テキスト (n=100) と NS テキスト (n=100) における SEM の頻度に対数尤度比検定 (LLR) を実行し，有意であった語をまとめたものである。なお，LLR は，検定によって導出された統計量を示し，NNS による過剰使用の場合には正の値，過少使用の場合には負の値を与えている。

表 4: 有意な SEM

	LLR	
I	2079.87	***
we	1185.84	***
my	137.99	***
me	41.96	***

NNS は，*I*，*we*，*my*，*me* という 4 つの SEM を過剰使用している。Hyland (2001) が指摘しているように，SEM は，書き手の存在を前景化する強力な修辞的戦略である。しかしながら，論説文においては，書き手の存在はテキストの背後にあり，それ故，書き手がテキストの前面に出てきたときに修辞的効果が生まれるのである。言い換えれば，基本的に「客観的」なトーンを持ったテキストに突如「主観的」なトーンが現れるからこそ，そこが強調されるのである。だが， NNS のテキストでは，基本的なトーンが「主観的」であるため，SEM の修辞的効果は生まれていない。むしろ，書き手の存在が常に前面に出ているために，論説文に必要とされる客観性が殆ど見られない。

3.4.2 Boosters (BOO)

BOO とは，自らの意見と対立する意見を遮断し，命題に対する確信度を強調する修辞法である。表 5 は，NNS テキストと NS テキストにおける BOO の頻度に対数尤度比検定を実行し，有意であった語句をまとめたものである。

表 5: 有意な BOO

	LLR	
think / thinks / thought	551.42	***
know / known	68.38	***
of course	41.90	***
believe / believed / believes	-35.61	***
show / showed / shown / shows	-29.92	***
must	22.75	***
establish / established	-10.36	**
prove / proved / proves	-8.72	**
certain / certainly	-5.60	*
truly / true	-4.68	*
realize / realized / realizes	-4.50	*

表 5 を見ると，NNS は，*think* を過剰使用している一方で，NS は *believe* や *realize* のようなより確信度の高い動詞を好み，*prove* や *show* で情報が「事実」であると強調している。また，*certainly* や *truly* のような *-ly* 副詞を巧みに操ることで，自らの論理展開をサポートしている。そして，NNS が確信を強調するときは，*must* と *of course* を多用する。

日本人学習者による *think* という語，とりわけ *I think* というコロケーションの過剰使用は，日本語の「思う」の影響である。ただ，和英辞典には，「思う」に対応する英語として，*think* 以外に，*consider*，*believe*，*expect*，*feel*，*wish*，*wonder*，*suspect*，*imagine*，*suppose*，*guess* などを含む多くの表現が載っている（本稿末尾の付録を参照）。NNS は，NS のようにこれらの語を文脈に応じて使い分けることができていないことが *think* の過剰使用の原因となっている。

さらに，外山 (1986) が指摘しているように，日本人学習者が産出する *think* は，NS が BOO の意味で用いているのとは異なり，*it seems to me* のような HED に近い意味で用いている。

3.3.3 Hedges (HED)

HED とは，自らの意見と対立する意見を想定し，情報を「事実」(fact) としてではなく意見「意見」(opinion) として提示する修辞法のことである。表 6 は，NNS テキストと NS テキストにおける HED の頻度に対数尤度比検定を実行し，有意であった語をまとめたものである。

表 6: 有意な HED

	LLR	
claim / claimed / claims	-85.37	***
would / wouldn't	-82.05	***
argue / argued / argues	-39.29	***
maybe	28.72	***
almost	19.38	***
sometimes	9.74	**
about	9.53	**
suggest / suggested / suggests	-7.74	**
seems	-4.81	*
feel / feels / felt	-4.27	*

表 6 を見ると，NNS が過剰使用している HED は全て副詞 (*maybe, almost, sometimes, about*) である一方，NNS が過少使用している HED は全て動詞 (*claim/claimed/claims, argue/argued/argues, suggest/suggested/suggests, seems, feel/feels/felt*) と助動詞 (*would/wouldn't*) である。

NNS が *would* のような仮定標識の助動詞を過少使用することはこれまでも報告されてきたが (Hyland & Milton

1997), *maybe* や *almost* のような単純副詞の過剰使用をすることは非常に興味深い。NNS がこれらの表現を好む理由は，*it seems that* や *it would be* のように文全体の構造を意識しないと使えない表現とは異なり，単純に 1 語を挿入するだけで HED の意味合いが付加されるからである。

3.3.4 Frame markers (FRM)

FRM とは，談話の順序 (e.g. *finally, firstly*)，段階 (e.g. *in summary, now*), 目的 (e.g. *intention, purpose*), 転換 (e.g. *back to, shift to*) に言及するものである。表 7 は，NNS テキストと NS テキストにおける FRM の頻度に対数尤度比検定を実行し，有意であった語句をまとめたものである。

表 7: 有意な FRM

	LLR	
want to	211.27	***
now	86.20	***
second / secondly	23.34	***
first / firstly / first of all	17.97	***
in short	7.56	**

表 7 を見ると，表中の全ての表現を NNS が過剰使用している。その中でも，NNS は，*firstly* や *secondly* などの sequencing と呼ばれる表現を好んで使っている。このような sequencing の過剰使用は, 書き手がそれまで受けてきた作文指導の影響と思われるが, このような接続語句の過剰使用は“artificial, mechanical prose” (Zamel 1983)であるという印象を読み手に与えかねない。

4. おわりに

本研究では，談話分析の観点から metadiscourse markers を説明変数とし，回帰木を用いて，NNS テキストと NS テキストの分類実験を行なった。その結果，92.0%の精度で両者を正しく分類することができた。また，SEM，BOO，HED，FRM が NNS テキストと NS テキストを識別する指標であることが判明した。

今後の課題としては，単なる統計的分析にとどまらず，本研究から得られた結果を実際の教室指導や教材作成に生かし，どうすれば「英語らしい」文章が書けるようになるのかという点を模索していかなければならない。

註

本研究の一部は，第 22 回「英検」研究助成（研究部門）「テキストマイニングによる学習者作文における談話能力の測定と評価」によって行なわれたものである。

参考文献

Hyland, K. (2001). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics, 34,* 1091-1112.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing.* New York: Continuum.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6,* 183-205.

杉浦正利 (2004).『なぜ英語母語話者は英語学習者が話すのを聞いてすぐに母語話者でないとわかるの』平成 13 年度～15 年度科学研究費補助金 基盤研究(C) 研究成果報告書.

田中省作・藤井宏・冨浦洋一・徳見道夫 (2006).「NS/NNS 論文分類モデルに基づく日本人英語科学作文の特徴抽出」『英語コーパス研究』*13,* 75-87.

外山滋比古 (1986).「『思われる』と『考える』」『思考の整理学』東京: ちくま文庫.

Zamel, V. (1983). Teaching those missing links in writing. *ELT Journal, 37,* 22-29.

付録: goo 和英辞典の「思う」の項

*《考える》think ((of, about)); 《見なす》consider [regard, look upon] ((as)); 《信じる》believe; 《予期する》expect; 《感じる》feel; 《願望》wish [desire, want]; 《いぶかる》wonder; 《ではないかと思う》suspect; 《想像する》imagine; 《推測する》suppose; guess; 《愛する》love; care for. **o** …しようとー be going to ((do)); intend ((to do)). **o** 良く［悪く］ー think well [ill] of. **o** 工事の進展がーに任せない The construction is not progressing as we expected. **o** 仕事がーように行かない Work isn't going as well as I would like